Arquitetura e pipeline de automatização da governança do grafo de conhecimento da rede ELLAS

Rodgers Fritoli
rodfritoli@hotmail.com
https://orcid.org/0009-0001-7696-7600
Universidade Tecnológica Federal do Paraná, Brasil

Rita Cristina Galarraga Berardi
ritaberardi@utfpr.edu.br
https://orcid.org/0000-0002-0281-8952
Universidade Tecnológica Federal do Paraná, Brasil

Recibido: 23 de agosto del 2023 / Aceptado: 4 de octubre del 2023
doi: https://doi.org/10.26439/interfases2023.n018.6623

RESUMO. Neste artigo é descrita uma proposta de arquitetura de integração de grafos de conhecimento e um pipeline que automatiza a transformação de dados em planilhas em grafos de conhecimento em triplas RDF, para a plataforma de dados abertos da rede de pesquisa ELLAS. O objetivo principal do projeto da rede pesquisa ELLAS é gerar dados comparáveis sobre as lacunas de gênero em STEM na América Latina, e analisar como os dados possuem diferentes fontes e formatos. Para a construção será utilizada web semântica, adicionando significados aos dados. A proposta da arquitetura e do pipeline é para realizar atividades de Extração, Transformação e Carregamento para o armazenamento e disponibilização dos dados para usuários finais.

PALAVRAS CHAVE: arquitetura de grafo de conhecimento / pipeline

ARQUITECTURA Y PIPELINE PARA AUTOMATIZAR EL GOBIERNO DEL GRAFO DE CONOCIMIENTO DE LA RED ELLAS

RESUMEN. Este artículo describe una propuesta de arquitectura de integración de gráficos de conocimiento y un *pipeline* que automatiza la transformación de datos de hojas de cálculo en gráficos de conocimiento en tripletas RDF para la plataforma de datos abiertos de la red de investigación ELLAS. El objetivo principal del proyecto de la red de investigación ELLAS es generar datos comparables sobre brechas de género en STEM en América Latina, y como los datos tienen diferentes fuentes y formatos para su construcción, se utilizará la web semántica, agregando significados a los

datos. La propuesta de arquitectura y pipeline es realizar actividades de (Extracción, Transformación y Carga) para almacenar y poner los datos a disposición de los usuarios finales.

PALABRAS CLAVE: arquitectura de gráfico de conocimiento / *pipeline*

ARCHITECTURE AND PIPELINE FOR AUTOMATING ELLAS NETWORK KNOWLEDGE NETWORK GOVERNANCE

ABSTRACT. In this article, a proposed architecture for integrating knowledge graphs and a pipeline that automates the transformation of spreadsheet data into RDF triple knowledge graphs for the open data platform of the ELLAS research network is described. The main goal of the ELLAS research network project is to generate comparable data on gender gaps in STEM in Latin America, and since the data comes from various sources and formats, semantic web will be used to add meaning to the data. The architecture and pipeline proposal is to perform Extract, Transform, and Load (ETL) activities for data storage and availability to end users.

KEYWORDS: knowledge graph architecture / pipeline

1. INTRODUÇÃO

A baixa representatividade de mulheres na área de STEM (Ciência, Tecnologia, Engenharia, Artes e Matemática) é um problema complexo, pois as mulheres precisam enfrentar desafios sociais e culturais para alcançar um cargo de liderança ou seguir na área de STEM (Branisa et al., 2021; Guzman et al., 2020; Hyvönen, 2020). A compreensão desse cenário exige a análise de diversas variáveis, como: dados demográficos, cultura, religião, dados governamentais, pois seria muito superficial afirmar, por exemplo, que somente estereótipos de gênero ou falta de modelos femininos de sucesso possam desencorajar as mulheres a seguirem em STEM, a questão vai além. A presença das mulheres tem aumentado na ciência de uma forma global, porém existe uma representação baixa nas áreas de conhecimento de ciências exatas e engenharias. Uma iniciativa do governo brasileiro, através de uma entidade ligada ao Ministério da Ciência, Tecnologia e Inovações para incentivo à pesquisa no Brasil (CNPq), foi criar o programa Mulher e Ciência. Espera-se que ações como essas possam se transformar em uma política pública que visa a reduzir a segregação horizontal e vertical das mulheres nas áreas STEM (Graphdb Ontotext, 2023).

A análise dos motivos da baixa representatividade requer a compreensão de dados provenientes de várias fontes e formatos, pois atualmente não existe uma fonte de informação única que abranja completamente esse tema, e de uma forma centralizada, o cenário atual também não é preciso. Neste domínio, é mais comum encontrar dados presentes em diversos artigos, porém os pesquisadores, quando realizam o levantamento, de modo geral, não deixam os dados disponíveis para reaproveitamento por outros pesquisadores de forma estruturada. Outro problema é que mesmo em plataformas existentes, há falta sobre dados no contexto latino, a exemplo, temos a plataforma da UNESCO que possui em suas bases de dados informações relacionadas à causa, porém não existem dados da América Latina.

Outro cenário é a devida filtragem de dados relacionados à causa, em bases que não possuem este fim, mas que podem contribuir com a discussão, como é o caso das bases do Instituto Nacional de Estudos e Pesquisas Educacionais (INEP)[1] que disponibiliza seus dados em diversas planilhas, tornando difícil a busca de informação sem que haja um pré-processamento dos dados. Por fm, as informações não chegam para quem tem poder de tomada de decisão, por exemplo pessoas responsáveis por criar políticas públicas para discutir o problema, pois o espalhamento, a não padronização e a falta de fontes centralizadas torna a tarefa de realizar análises precisas e identificar padrões consistentes bastante onerosa e difícil.

1 https://www.gov.br/inep/pt-br

A plataforma desenvolvida pela rede ELLAS[2] de pesquisa tem como objetivo providenciar uma maneira de centralizar esses dados que não estão conectados e, assim, favorecer que pesquisadores produzam estudos para recomendações de políticas para instituições públicas e privadas. Tal produção está concentrada em três países da América do Sul (Brasil, Bolívia e Peru) e pretende abordar sistematicamente a questão da liderança feminina em STEM revisando a literatura internacional. A ideia é criar uma plataforma de dados abertos e conectados por ontologias (assegurando sua organização, acessibilidade e reutilização) (de Araújo e Tonini, 2020), com o intuito de simplificar e incentivar o uso por qualquer indivíduo interessado no tema.

Para esse tema ainda não existe uma arquitetura bem definida que organize e conecte todos os dados produzidos por diferentes equipes, com uma curadoria dos dados desde o recebimento em sua forma mais crua até o processo de transformação dos dados em triplas e conexão, automatização das cargas de dados e condução dos dados até o processo de disponibilização dos dados. A arquitetura proposta não terá a responsabilidade de efetuar o tratamento dos dados em sua forma bruta, uma vez que esse procedimento será conduzido pelas equipes multidisciplinares. O fluxo de dados direciona as informações para o estágio de carregamento, onde a ontologia será aplicada, os dados serão convertidos em triplas e, por último, a disponibilização do acesso será realizada.

Os dados não seguirem um padrão de colunas é um fator dificultador e torna difícil uma abordagem para extrair, transformar e carregar (ETL). A utilização de um banco de dados relacional para este tipo de projeto não é recomendada, por conta disto foi previamente escolhido um banco de dados que comporte web semântica e a carga de ontologias, além disso, o banco escolhido possui diversas formas de componentes de conexão que permite a conectividade perfeita e simplifica a construção do grafo de conhecimento.

Uma das decisões quanto à arquitetura e governança diz respeito à adoção de um servidor em uma nuvem, o que pode escalar de uma forma não esperada o consumo de recurso financeiro, para tal monitorar será necessário criar relatórios de acompanhamento de consumo de processamento e memória. O objetivo é construir uma arquitetura para um processo de conexão automatizado, conduzindo os dados desde o seu formato original até a sua integração no grafo de conhecimento semântico, que será utilizado o triplestore GraphDB a infraestrutura de cloud computing escolhida será a AWS.

2. ARQUITETURA DO SISTEMA

Até o momento foi realizado um piloto de configuração de um servidor na AWS para começar a realizar alguns testes com o pipeline a ser desenvolvido. Para a instalação do

2 https://ellas.ufmt.br/pt/inicio/

GraphDB e do Pentaho Data Integration foi criado um servidor na AWS. Este servidor foi configurado com uma capacidade básica de 4 GB de RAM e está utilizando o sistema operacional Windows Server 2022. Foi instalada a versão 10.2 do GraphDB Desktop. Para permitir o acesso externo ao servidor, foi necessário liberar as portas 7200 e 7333 no firewall. Essa configuração permite que as conexões sejam estabelecidas com o GraphDB e o Pentaho Data Integration a partir de fontes externas, permitindo o gerenciamento destas plataformas, Na Figura 1 é possível visualizar de forma sintética como está projetada a proposta da Arquitetura e do Pipeline.

O GraphDB é um banco de dados orientado a grafos, ou um triplestore, que permite armazenar dados complexos e altamente conectados. É especialmente adequado para a utilização de web semântica para aplicações como redes sociais, informações geoespaciais, Ontologias e Grafos de conhecimento (W3C, 2023b). O GraphDB suporta diversos formatos, incluindo RDF, OWL, JSON e CSV, facilitando a importação e exportação de dados.

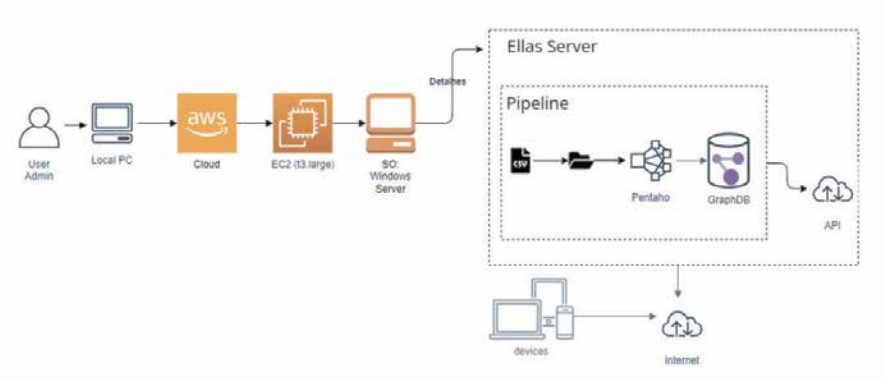
Além da definição da arquitetura e de ferramentas de armazenamento e processamento, é necessário projetar um Pipeline que prevê a manipulação dos dados desde a sua forma crua até a forma final a ser armazenada e disponibilizada. Um dos dados presentes na plataforma é coletado pelos especialistas do domínio mulheres em STEM, que coletaram dados da literatura e estruturaram em planilhas. Para incluí-los no banco de dados é necessário instanciar a ontologia, por meio de um processo de triplificação que é a conversão de dados em triplas. A triplificação é um processo que transforma os dados em um formato padrão para triplas RDF, seguindo o modelo de sujeito-predicado-objeto (W3C, 2023a). No entanto, considerando a variabilidade das bases de dados e visando uma escalabilidade da plataforma, é necessário adotar uma abordagem de modelagem de dados e realizar uma etapa de preparação dos dados antes da triplificação.

A modelagem de dados envolve a definição dos conceitos, propriedades e relações relevantes para representar os dados de forma semântica, alinhada à ontologia desenvolvida pela rede ELLAS de pesquisa. Essa etapa é importante para garantir a correta representação dos dados e permitir a realização de consultas e inferências significativas.

Na Figura 1, observa-se uma versão macro da arquitetura composta pelo método de acesso pelo cloud computing AWS. Através de uma estação local do administrador da arquitetura, o serviço da AWS, chamado EC2, que é a plataforma em que o usuário pode criar máquinas virtuais, o desenho macro do pipeline orquestrado pelo Pentaho Data Integration e a representação da saída para os usuários consumidores

Figura 1

Versão Inicial e compacta da Arquitetura do Sistema

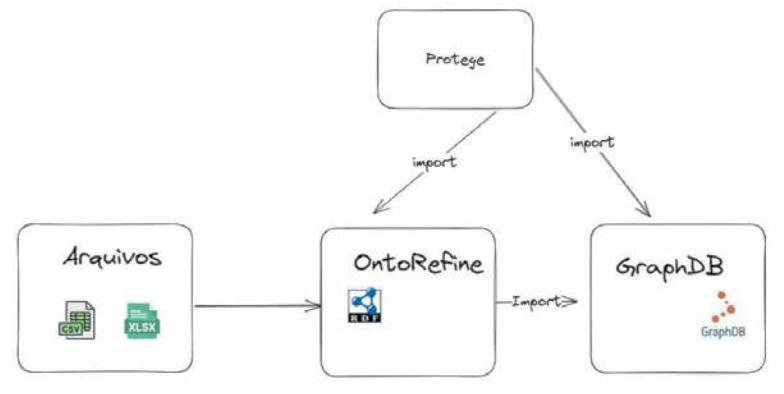


3. PIPELINE

Inicialmente, o processo de importação de dados é realizado de forma manual, envolvendo a leitura de arquivos em formatos de planilha. Em seguida, esses dados são processados no Ontotext Refine[3], em que são realizadas etapas de limpeza e mapeamento. Posteriormente, os dados são importados manualmente no formato RDF para o banco GraphDB. O processo de carga manual está ilustrado na Figura 2.

Figura 2

Pipeline manual



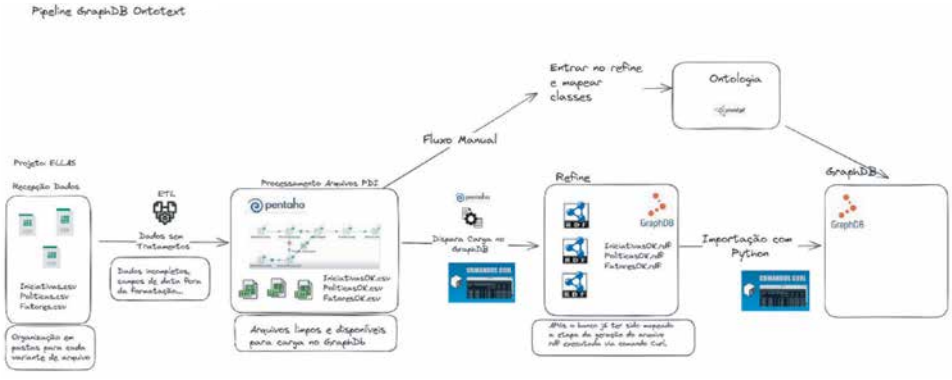
O cenário manual foi útil para criar e testar o processo, no entanto, pode ser trabalhoso e propenso a erros, demandando esforços consideráveis por parte dos usuários, além de não ser escalável e sustentável a longo prazo. Para melhorar a eficiência do

3 https://www.ontotext.com/products/ontotext-refine/

processo de carga de dados, desenvolveu-se uma abordagem automatizada, utilizando o Pentaho Data Integration como orquestrador. Na Figura 3 é possível observar que para automatizar o processo manual foi necessário a definição dos seguintes passos: Definir a ferramenta de orquestração; realizar o processo de processamento do Ontotext Refine via comando; exportar os arquivos em formato RDF e gravar no banco de dados, todas as etapas sem intervenção manual.

Figura 3

Pipeline automatizado



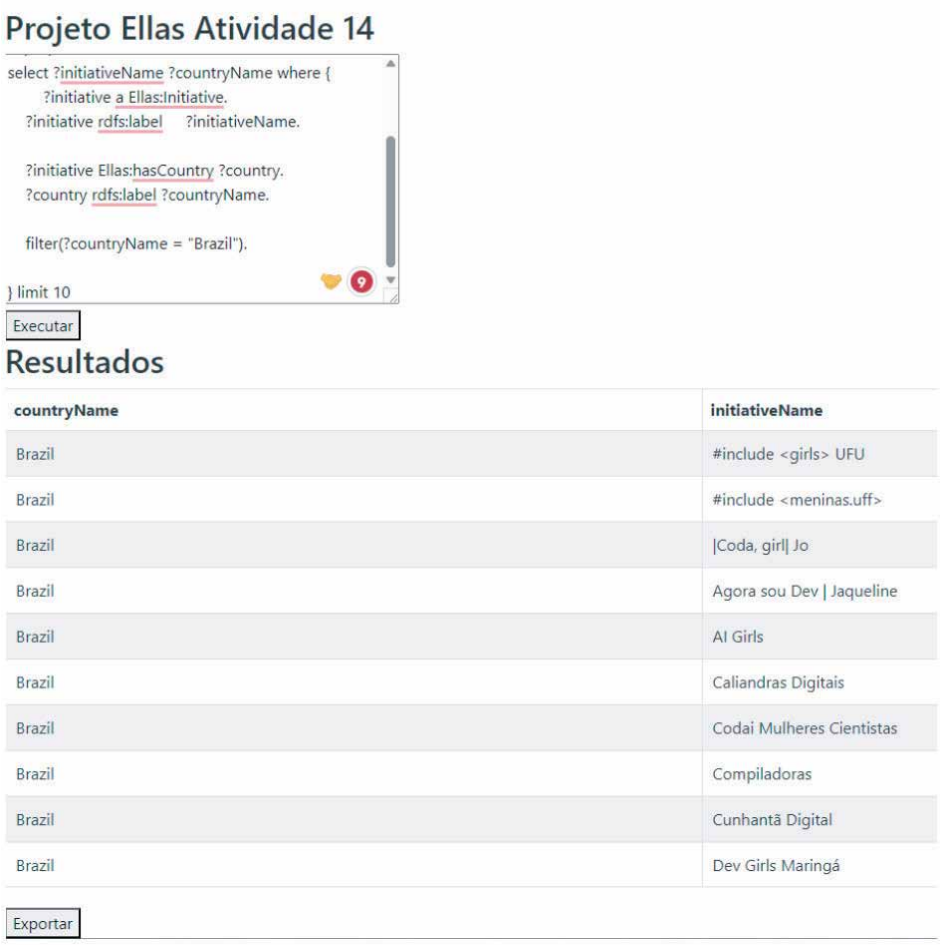
4. ACESSO À PLATAFORMA POR MEIO DO ENDPOINT SPARQL

Finalmente, para o acesso aos dados é necessário disponibilizar os dados para consultas, sem uma preocupação com a usabilidade para usuários e sem conhecimento técnico. No contexto em que o GraphDB não possui um endpoint genérico para consultas SPARQL, desenvolveu-se na configuração piloto do servidor uma página HTML para suprir essa lacuna, permitindo execuções diretas de consultas. Enquanto muitas abordagens exigem ferramentas externas, a escolha aqui foi integrar-se diretamente ao GraphDB via HTML personalizado, simplificando o processo. A solução, ilustrada na Figura 4, oferece aos usuários mais técnicos uma interface acessível para consultas SPARQL, beneficiando especialmente aqueles usuário já familiarizados com a linguagem SPARQL e possibilita uma conexão com os dados em aplicações que sejam desenvolvidas com base nos dados da rede ELLAS. Para a implementação, a linguagem de programação Python foi utilizada.

No Python, foi empregada a biblioteca rdflib, que permite a leitura e conexão com o GraphDB. Para a criação da página web, foi utilizado o framework Flask para renderizar o template HTML. Essa abordagem permitiu criar, Figura 4, uma solução customizada para a execução de consultas SPARQL no GraphDB, integrando a biblioteca rdflib e o framework Flask para fornecer uma interface simples ao usuário na página HTML.

Figura 4

Endpoint de consulta Sparql

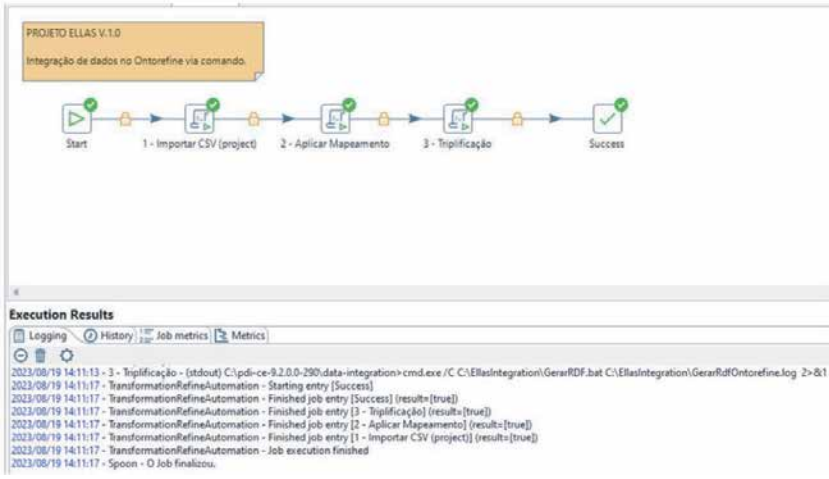


5. RESULTADOS INICIAIS

O projeto está em desenvolvimento e a proposta de integração automatizada foi simulada com uma amostra de 300 linhas de dados extraídos da literatura e estruturados pelos especialistas de domínio. Na integração dos arquivos do projeto foi utilizado o Pentaho Data Integration, aproveitou-se o mapeamento e a ontologia previamente estabelecidos para esta tarefa. O Pentaho localiza os arquivos no diretório especificado, inicia um projeto no Ontotext Refine através de um comando e aplica o mapeamento aos dados tabulares. Isso prepara os dados para a conversão em triplas RDF (Resource Description Framework), o formato compatível com o GraphDB. Após a geração do arquivo RDF, os dados são registrados com sucesso no banco de dados. O processo de orquestração pode ser visto na figura 5 e o exemplo da execução.

Figura 5

Transformação de dados e inserção de dados com Pentaho



O uso deste piloto inicial mostrou benefícios e que as decisões quanto às ferramentas estão adequadas para o projeto. O desenvolvimento de um pipeline automatizado executou com excelência os mesmos passos desenvolvidos na versão manual do processo, o que mostra que a arquitetura satisfaz as necessidades iniciais para o desenvolvimento da plataforma. Com a solução em uso, vislumbra-se potenciais melhorias. A integração da biblioteca rdflib ao framework Flask pode ser ampliada para funcionalidades adicionais e melhor experiência na página HTML.

6. CONCLUSÕES

Foi descrita, neste artigo, uma arquitetura com pipeline para automatização de processo de conversão de dados em planilhas para grafos de conhecimento RDF. Foi realizada uma simulação completa da Arquitetura, orquestrando a integração de dados e otimizando a gestão e consulta de dados. Com a infraestrutura proposta estabelecida, torna-se possível criar interfaces de consulta eficientes e soluções de integração de dados personalizada.

No entanto, a atual arquitetura não prevê estratégias avançadas de clusterização e redundância de dados. Se a plataforma exigir elementos de alta disponibilidade, escalabilidade e soluções é indicado o Kafka, conhecido por sua capacidade robusta de processamento e cargas de dados, que poderia ser considerado para aprimorar a arquitetura proposta. Assim, como proposta para futuras pesquisas, recomenda-se a investigação de mecanismos de clusterização e a integração de ferramentas como o Kafka para fortalecer e expandir a capacidade da infraestrutura proposta. Uma perspectiva interessante é a integração com modelos de linguagem, como o ChatGPT,

proporcionando uma interface conversacional e facilitando consultas naturais ao GraphDB.

Com a adição de uma interface, as consultas aos grafos podem se tornar mais intuitivas, os usuários não precisariam saber a sintaxe específica da consulta SPARQL ou teriam a possibilidade de fazer perguntas em linguagem natural. Outro ponto importante é integração dos grafos com componentes de visualização de dados.

7. REFERÊNCIAS

Branisa, B., Cabero, P., & Guzman, I. (2021). *The main factors explaining IT Career Choices of Female Students in Bolivia*. AMCIS 2021 Proceedings.

De Araújo, M. T., & Tonini, A. M. (2020). A PARTICIPAÇÃO DAS MULHERES NAS ÁREAS DE STEM (SCIENCE, TECHNOLOGY ENGINEERING AND MATHEMATICS). *Revista de Ensino de Engenharia*, *38*(3). http://revista.educacao.ws/revista/index.php/abenge/article/view/1693/905

Graphdb Ontotext. (2023). *What is GraphDB?* https://graphdb.ontotext.com/documentation/

Guzman, I., Berardi, R., Maciel, C., Cabero Tapia, P., Marin-Raventos, G., Rodriguez, N., & Rodriguez, M. (2020). *Gender Gap in IT in Latin America*. AMCIS 2020 Proceedings.

Hyvönen, E. (2020). *Linked Open Data Infrastructure for Digital Humanities in Finland.* CEUR Workshop Proceedings. http://urn.fi/URN:NBN:fi:aalto-202101251587

W3C. (2023a). *Resource Description Framework (RDF)*. https://www.w3.org/RDF/

W3C. (2023b). *Web Semântica*. https://www.w3.org/standards/semanticweb/